Online Localization and Tracking of Multiple Moving Speakers in Reverberant Environments

Xiaofei Li*, Yutong Ban*, Laurent Girin, Xavier Alameda-Pineda and Radu Horaud

***Abstract*—We address the problem of online localization and tracking of multiple moving speakers in reverberant environments. The paper has the following contributions. We use the direct-path relative transfer function (DP-RTF), an inter-channel feature that encodes acoustic information robust against reverberation, and we propose an online algorithm well suited for estimating DP-RTFs associated with moving audio sources. Another crucial ingredient of the proposed method is its ability to properly assign DP-RTFs to audio-source directions. Towards this goal, we adopt a maximum-likelihood formulation and we propose to use exponentiated gradient (EG) to efficiently update source-direction estimates starting from their currently available values. The problem of multiple speaker tracking is computationally intractable because the number of possible associations between observed source directions and physical speakers grows exponentially with time. We adopt a Bayesian framework and we propose a variational approximation of the posterior filtering distribution associated with multiple speaker tracking, as well as an efficient variational expectation maximization (VEM) solver. The proposed online localization and tracking method is thoroughly evaluated using two datasets that contain recordings performed in real environments.**

***Index Terms*—Inter-channel acoustic features, reverberant environments, sound-source localization, multiple target tracking, speaker tracking, Bayesian variational inference, expectation-maximization.**

I. Introduction

The localization and tracking of multiple speakers in real world environments are very challenging tasks, in particular in the presence of reverberation and ambient noise and of natural conversations, e.g. short sentences, speech pauses and frequent speech turns among speakers. Methods based on time differences of arrival (TDOAs) between microphones, such as generalized cross-correlation [1], are typically used for single-speaker localization, e.g. [2]. In the case of multiple speakers, beamforming-based methods, e.g. steered-response power (SRP) [3], and subspace methods, e.g. multiple signal classification (MUSIC) [4], are widely used. The W-disjoint orthogonality (WDO) principle [5] assumes that the audio signal is dominated by a single audio source in small regions of the time-frequency (TF) domain. This assumption is particularly valid in the case of speech signals. Applying the short-time Fourier transform (STFT), or any other TF representation, inter-channel localization features, such as the interaural phase differences (IPDs) [5], can be extracted. In [5], multiple-speaker localization is based on the histogram of inter-channel features, which is suitable only in the case where there is no wrapping of phase measures. In [6], a Gaussian mixture model (GMM) is used as a generative model of the inter-channel features of multiple speakers, with each GMM representing one speaker, and each GMM component representing one candidate inter-channel time delay. An expectation maximization (EM) algorithm iteratively estimates the component weights and assigns the features to their corresponding candidate time delays. This method overcomes the phase ambiguity problem by jointly considering all frequencies in the likelihood maximization procedure. After maximizing the likelihood, the azimuth of each speaker is given by the component that has the highest weight in the corresponding GMM. The complex-valued version of IPD, i.e. the pair-wise relative phase ratio (PRP), is used in [7]. Instead of setting one GMM for each speaker, a single complex Gaussian mixture model (CGMM) is used for all speakers with each component representing one candidate speaker location. After maximizing the likelihood of the PRP features, with an EM algorithm, the weight of each component represents the probability that there is an active speaker at the corresponding candidate location. Therefore, for an unknown number of speakers, counting and localization of active speakers can be jointly carried out by selecting components with large weights.

The inter-channel features and associated localization methods mentioned above assume a direct-path propagation model: hence, they perform poorly in reverberant environments. To overcome this limitation, several TDOA estimators based on system identification were proposed in [8]–[11]. In [12] it is proposed to use the DP-RTF as a TF-domain inter-channel localization feature robust against reverberation. The estimation of the DP-RTF is based on the identification of the room impulse response (RIR) in the STFT-domain, i.e. the convolutive transfer function (CTF) [13], [14]. Overall, the method of [12] combines the merits of robust TDOA estimators [8]–[11] and of the WDO assumption mentioned above.

To localize moving speakers, one-stage methods such as SRP and MUSIC can be directly used using frame-wise spatial spectrum estimators. In contrast, methods based on inter-channel features require to assign frame-wise features to speakers in an adaptive/recursive way, e.g. the smoothed histogram method of [15]. Similar to [7], [16] uses one CGMM for each predefined speaker; the model is plugged into a recursive EM (REM) algorithm in order to update the

* X. Li and Y. Ban have equally contributed to the paper.

X. Li, Y. Ban, X. Alameda-Pineda and R. Horaud are with Inria Grenoble Rhône-Alpes and with Univ. Grenoble Alpes, France.

L. Girin is with Univ. Grenoble Alpes, Grenoble-INP, GIPSA-lab and Inria.

This work was supported by the ERC Advanced Grant VHIA #340113.

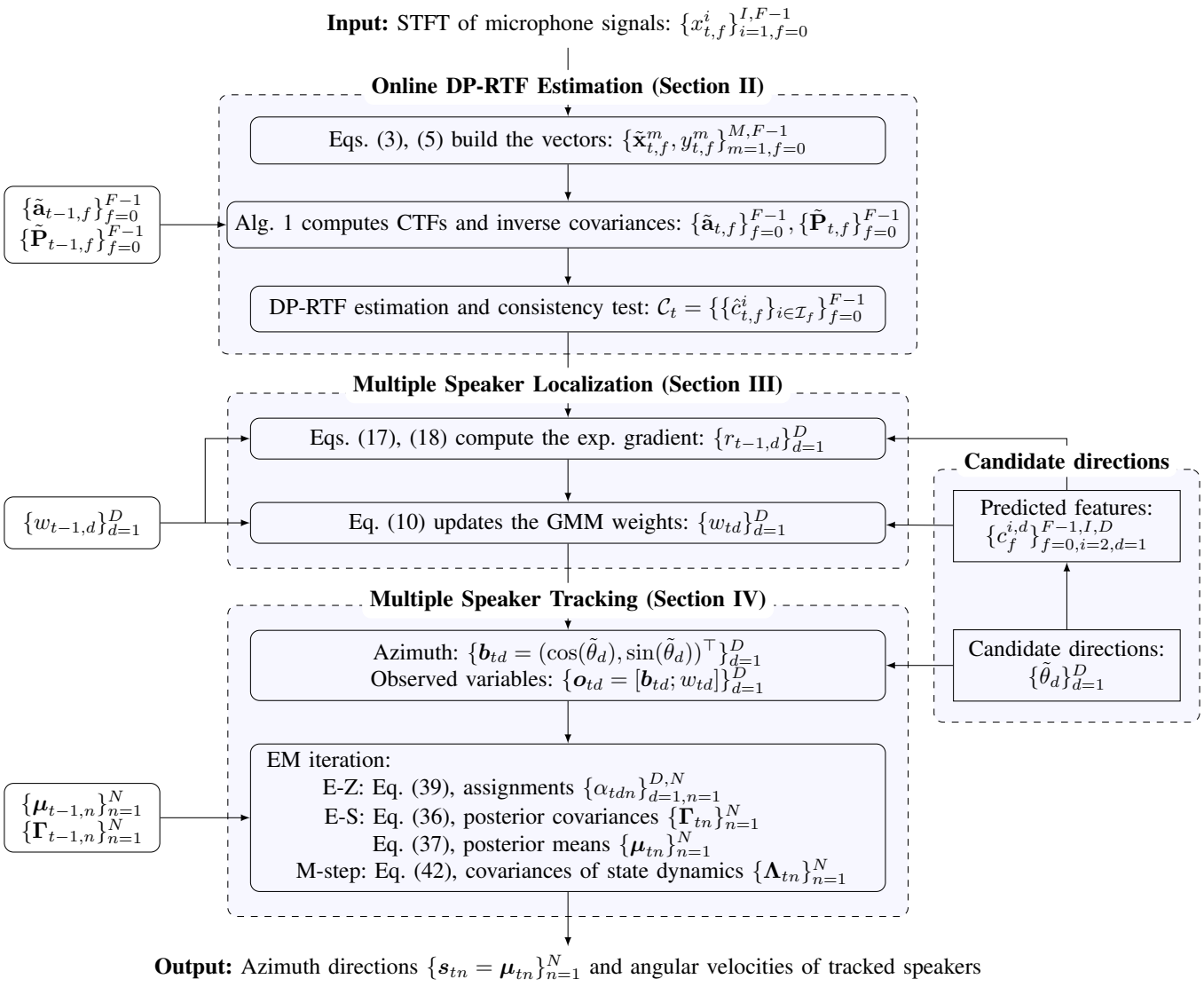


Fig. 1: Flowchart of the proposed multiple-speaker localization and tracking methodology.

mixture's weights.

Speaker tracking methods are generally based on Bayesian inference which combines localization with dynamic models in order to estimate the posterior probability distribution of audio-source directions, e.g. [17]–[19]. Kalman filtering and particle filtering were used in [20] and in [21], respectively, for tracking a single audio source. In order to address the problem of multiple speakers, possibly with unknown and time-varying number of speakers, additional discrete latent variables are needed, i.e. observation-to-speaker assignments, as well as speaker-birth and -death processes, e.g. [22], [23]. Sampling-based methods were widely used, e.g. extended particle filtering [24]–[26], or sequential Monte Carlo implementation of the probability hypothesis density (PHD) filter [27], [28]. However, the computational burden of sampling-based methods can be prohibitive in practice. Under some assumptions, the multiple-target tracking GMM-PHD filter of [29] has an analytical solution and is computationally efficient: it was adopted for multiple-speaker tracking in [18].

In this paper we propose a method for the simultaneous localization and tracking of multiple moving speakers (please refer to Figure 1 for a method overview). The paper has the following original contributions:

- Since we deal with moving speakers or, more generally, with moving audio sources, DP-RTF features are computed using the *online* CTF estimation framework presented in [30], based on recursive least squares (RLS), rather than using the *batch* CTF estimation of [12] which assumes static audio sources. The online RLS algorithm has a faster convergence rate than the least mean squares (LMS) algorithms described in [8], [9]. This is important when dealing with multiple moving sources, where the adaptive estimator is required to quickly switch between multiple sources and to deal with moving sources.

- A crucial ingredient of multiple speaker localization is to properly assign acoustic features, i.e. DP-RTFs, to audio-source directions. We adopt the maximum-likelihood formulation of [7]. We propose to use EG [31] to update the source directions from their current estimated values. The EG-based recursive estimator proposed below is better suited for moving sources/speakers than the batch estimator proposed in [12].
- The problem of multiple speaker tracking is computationally intractable because the number of possible associations between acoustic features and sources/speakers grows exponentially with time. In this paper we adopt a Bayesian variational approximation of the posterior filtering distribution which leads to an efficient VEM algorithm. In order to deal with a varying number of speakers, we propose a birth process which allows to initialize new speakers at any time.

This paper is an extended version of [30] which has proposed an online DP-RTF method that has been combined with REM to estimate the source directions. In this paper, while we keep the DP-RTF method of [30] we propose to use EG. The advantages of using EG instead of REM are described in detail in Section III. Moreover, the multiple speaker tracking method is completely novel.

The paper is organized as follows (please refer to Figure 1). Section II presents the online DP-RTF estimation method. Section III describes the EG-based speaker localization method and Section IV describes the variational approximation of the tracker and the associated VEM algorithm. Section V presents an empirical evaluation of the method based on experiments performed with real audio recordings. Section VI concludes the paper. Supplemental materials are available on our website.[1]

II. Recursive Multichannel DP-RTF estimation

A. Recursive Least Squares

For the sake of clarity, we first consider the noise-free single-speaker case. In the time domain $x^i(\tau) = a^i(\tau) \star s(\tau)$ is the i-th microphone signal, $i = 1, \dots, I$, where τ is the time index, $s(\tau)$ is the source signal, $a^i(\tau)$ is the RIR from the source to the i-th microphone, and $\star$ denotes the convolution. Applying the STFT and using the CTF approximation, for each frequency index $f = 0, \dots, F-1$ we have:

$$x^i_{t,f} = a^i_{t,f} \star s_{t,f} = \sum_{q=0}^{Q-1} a^i_{q,f} s_{t-q,f}, \tag{1}$$

where $x^i_{t,f}$ and $s_{t,f}$ are the STFT coefficients of the corresponding signals, and the CTF $a^i_{t,f}$ is a sub-band representation of $a^i(\tau)$. Here, the convolution is executed with respect to the frame index t. The number of CTF coefficients Q is related to the reverberation time of the RIR. The first CTF coefficient $a^i_{0,f}$ mainly consists of the direct-path information, thence the DP-RTF is defined as the ratio between the first CTF coefficients of two channels: $a^i_{0,f}/a^r_{0,f}$, where channel r is the reference channel.

Based on the cross-relation method [32], using the CTF model of one microphone pair (i, j), we have: $x^i_{t,f} \star a^j_{t,f} = x^j_{t,f} \star a^i_{t,f}$. This can be written in vector form as:

$$\mathbf{x}^{i\top}_{t,f} \mathbf{a}^j_f = \mathbf{x}^{j\ \top}_{t,f} \mathbf{a}^i_f, \tag{2}$$

with $\mathbf{a}^i_f = (a^i_{0,f}, \dots, a^i_{Q-1,f})^\top$, where $^\top$ denotes matrix/vector transpose, and $\mathbf{x}^i_{t,f} = (x^i_{t,f}, \dots, x^i_{t-Q+1,f})^\top$. The CTF vector involving all channels is defined as $\mathbf{a}_f = (\mathbf{a}^{1\top}_f, \dots, \mathbf{a}^{I\top}_f)^\top$. There is a total of $I(I-1)/2$ distinct microphone pairs, indexed by (i, j) with $i = 1, \dots, I-1$ and $j = i+1, \dots, I$. For each pair, we construct a cross-relation equation in terms of $\mathbf{a}_f$. For this aim, we define:

$$\mathbf{x}^{ij}_{t,f} = [\underbrace{0, \dots, 0}_{(i-1)Q}, \mathbf{x}^{j\ \top}_{t,f}, \underbrace{0, \dots, 0}_{(j-i-1)Q}, -\mathbf{x}^{i\ \top}_{t,f}, \underbrace{0, \dots, 0}_{(I-j)Q}]^\top. \tag{3}$$

Then, for each pair (i, j), we have:

$$\mathbf{x}^{ij\ \top}_{t,f} \mathbf{a}_f = 0. \tag{4}$$

Let's assume, for simplicity, that the reference channel is $r = 1$. To avoid the trivial solution $\mathbf{a}_f = \mathbf{0}$ of (4), we constrain the first CTF coefficient of the reference channel to be equal to 1. This is done by dividing both sides of (4) by $a^1_{0,f}$ and by moving the first entry of $\mathbf{x}^{ij}_{t,f}$, denoted by $-y^{ij}_{t,f}$, to the right side of (4), which rewrites as:

$$\tilde{\mathbf{x}}^{ij\ \top}_{t,f} \tilde{\mathbf{a}}_f = y^{ij}_{t,f}, \tag{5}$$

where $\tilde{\mathbf{x}}^{ij}_{t,f}$ is $\mathbf{x}^{ij}_{t,f}$ with the first entry removed, and $\tilde{\mathbf{a}}_f$ is the relative CTF vector:

$$\tilde{\mathbf{a}}_f = \left(\frac{\tilde{\mathbf{a}}^{1\top}_f}{a^1_{0,f}}, \frac{\mathbf{a}^{2\top}_f}{a^1_{0,f}}, \dots, \frac{\mathbf{a}^{I\top}_f}{a^1_{0,f}} \right)^\top, \tag{6}$$

where $\tilde{\mathbf{a}}^1_f = (a^1_{1,f}, \dots, a^1_{Q-1,f})^\top$ denotes $\mathbf{a}^1_f$ with the first entry removed. For $i = 2, \dots, I$, the DP-RTFs appear in (6) as the first entries of $\frac{\mathbf{a}^{i\top}_f}{a^1_{0,f}}$. Therefore, the DP-RTF estimation amounts to solving (5).

Equation (5) is defined for one microphone pair and for one frame. In batch mode, the terms $\tilde{\mathbf{x}}^{ij\ \top}_{t,f}$ and $y^{ij}_{t,f}$ of this equation can be concatenated accross microphone pairs and frames to construct a least square formulation. For online estimation, we would like to update the $\tilde{\mathbf{a}}_f$ using the current frame t. For notational convenience, let $m = 1, \dots, M$ denote the index of a microphone pair, where $M = I(I-1)/2$. Then let the superscript ^{ij} be replaced with ^{m}. The fitting error of (5) is

$$e^m_{t,f} = y^m_{t,f} - \tilde{\mathbf{x}}^{m\ \top}_{t,f} \tilde{\mathbf{a}}_f. \tag{7}$$

At the current frame t, for the microphone pair m, RLS aims to minimize the error

$$J^m_{t,f} = \sum_{t'=1}^{t-1} \sum_{m'=1}^{M} \lambda^{t-t'} |e^{m'}_{t',f}|^2 + \sum_{m'=1}^{m} |e^{m'}_{t,f}|^2, \tag{8}$$

[1] https://team.inria.fr/perception/research/multi-speaker-tracking/

Algorithm 1 RLS at frame t

Input: $\tilde{\mathbf{x}}_{t,f}^m$, $y_{t,f}^m$, $m = 1, \dots, M$
Initialization: $\tilde{\mathbf{a}}_{t,f}^0 \leftarrow \tilde{\mathbf{a}}_{t-1,f}^M$, $\mathbf{P}_{t,f}^0 \leftarrow \lambda^{-1}\mathbf{P}_{t-1,f}^M$
for $m = 1$ to M **do**
$e_{t,f}^m = y_{t,f}^m - \tilde{\mathbf{x}}_{t,f}^{m\top}\tilde{\mathbf{a}}_{t,f}^{m-1}$
$\mathbf{g} = \mathbf{P}_{t,f}^{m-1}\tilde{\mathbf{x}}_{t,f}^{m*}/(1 + \tilde{\mathbf{x}}_{t,f}^{m\top}\mathbf{P}_{t,f}^{m-1}\mathbf{x}_{t,f}^{m*})$
$\mathbf{P}_{t,f}^m = \mathbf{P}_{t,f}^{m-1} - \mathbf{g}\tilde{\mathbf{x}}_{t,f}^{m\top}\mathbf{P}_{t,f}^{m-1}$
$\tilde{\mathbf{a}}_{t,f}^m = \tilde{\mathbf{a}}_{t,f}^{m-1} + e_{t,f}^m\mathbf{g}$
end for
Output: $\tilde{\mathbf{a}}_{t,f}^M$, $\mathbf{P}_{t,f}^M$

which sums up the fitting error of all the microphone pairs for the past frames and the microphone pairs up to m for the current frame. The forgetting factor $\lambda \in (0, 1]$ gives a lower weight to older frames, whereas all microphone pairs have the same weight at each frame. To minimize $J_{t,f}^m$, we set its complex derivative with respect to $\tilde{\mathbf{a}}_f^*$ to zero, where * denotes complex conjugate. We obtain an estimate of $\tilde{\mathbf{a}}_f$ at frame t for microphone pair m as:

$$\tilde{\mathbf{a}}_{t,f}^m = {\mathbf{R}_{t,f}^m}^{-1} r_{t,f}^m, \tag{9}$$

with

$$\mathbf{R}_{t,f}^m = \sum_{t'=1}^{t-1}\sum_{m'=1}^{M} \lambda^{t-t'} \tilde{\mathbf{x}}_{t',f}^{m'*}\tilde{\mathbf{x}}_{t',f}^{m'\top} + \sum_{m'=1}^{m} \tilde{\mathbf{x}}_{t,f}^{m'*}\tilde{\mathbf{x}}_{t,f}^{m'\top},$$

$$r_{t,f}^m = \sum_{t'=1}^{t-1}\sum_{m'=1}^{M} \lambda^{t-t'} \tilde{\mathbf{x}}_{t',f}^{m'*} y_{t',f}^{m'} + \sum_{m'=1}^{m} \tilde{\mathbf{x}}_{t,f}^{m'*} y_{t,f}^{m'}.$$

It can be seen that the covariance matrix $\mathbf{R}_{t,f}^m$ is computed based on the rank-one modification, thence its inverse, denoted by $\mathbf{P}_{t,f}^m$, can be computed using the Sherman-Morrison formula, without the need of matrix inverse. The recursion procedure is summarized in Algorithm 1, where $\mathbf{g}$ is the *gain vector*. The current frame t is initialized with the previous frame $t-1$. At the first frame, we initialize $\tilde{\mathbf{a}}_{1,f}^0$ as zero, and $\mathbf{P}_{1,f}^0$ as the identity. At each frame, all microphone pairs are related to the same CTF vector that corresponds to the current speaker direction, hence all microphone pairs should be simultaneously used to estimate the CTF vector of the current frame. In batch mode, this can be easily implemented by concatenating the microphone pairs. However, in RLS, to satisfy the rank-one modification of the covariance matrix, we need to process the microphone pairs one by one as shown in (8) and Algorithm 1. At the end of the iterations over all microphone pairs, $\tilde{\mathbf{a}}_{t,f}^M$ is the "final" CTF estimation for the current frame, and is used for speaker localization. The DP-RTF estimates, denoted as $\tilde{c}_{t,f}^i,\ i = 2, \dots, I$, are obtained from $\tilde{\mathbf{a}}_{t,f}^M$. Note that implicitly we have $\tilde{c}_{t,f}^1 = 1$.

B. Multiple Moving Speakers

So far, the proposed online DP-RTF estimation method has been presented in the noise-free single-speaker case. The noisy multiple-speaker case was considered in [12], but only for static speakers, i.e. batch mode, and in the two-channel case. We summarize the principles of this method and then explain in details the present online/multi-channel extension.

1) Estimation of the CTF vector: It is reasonable to assume that the CTF vector doesn't vary over a few consecutive frames and that only one speaker is active within a small region in the TF domain, due to the sparse representation of speech in this domain. Consequently, the CTF vector can be estimated over the current frame and a few past frames. An estimated CTF value, at each TF bin, is then assumed to be associated with only one speaker. The CTF vector computation in the case of multiple speakers can be carried out using the RLS algorithm, presented in Section II-A, by adjusting the forgetting factor λ to yield a short memory.

The forgetting factor λ is set to $\lambda = \frac{P-1}{P+1}$, where P is the number of frames being used. To efficiently estimate the CTF vector $\tilde{\mathbf{a}}_{t,f}^M$ of length $IQ-1$, we need $\rho \times (IQ-1)$ equations, where the parameter ρ should be chosen in such a way to achieve a good tradeoff between the validity of the above assumptions and robust estimation of $\tilde{\mathbf{a}}_{t,f}^M$. To guarantee that $\rho \times (IQ-1)$ equations are available, we need $P = \frac{\rho(IQ-1)}{I(I-1)/2} \approx \rho\frac{2Q}{I-1}$ frames. One may observe that the number of frames needed to estimate $\tilde{\mathbf{a}}_{t,f}^M$ decreases as the number of microphones increases.

2) Noise reduction: When noise is present, especially if the noise sources are temporally/spatially correlated, the CTF estimate can be contaminated. In addition, even in a low-noise case, many TF bins are dominated by noise due to the sparsity of speech spectra. To classify the speech frames and noise frames, and to remove the noise, we use the inter-frame spectral subtraction algorithm proposed in [30], [33].

The cross- and auto-power spectral density (PSD) between the convolution vector of the microphone signals, i.e. $\mathbf{x}_{t,f}^i$, and the current frame of the reference channel, i.e. $x_{t,f}^1$, is computed by averaging the cross- and auto-periodograms over frames. In the present work, we use recursive averaging:

$$\boldsymbol{\phi}_{t,f}^i = \beta\boldsymbol{\phi}_{t-1,f}^i + (1-\beta)\mathbf{x}_{t,f}^i x_{t,f}^{1\,*}, \quad i = 1, \dots, I, \tag{10}$$

where the smoothing parameter β is set to achieve a good tradeoff between low noise PSD variance and fast tracking of speech variation. The noise frames and speech frames are classified based on the minimum statistics [33] of the PSD of $x_{t,f}^1$, i.e. the first entry of $\boldsymbol{\phi}_{t,f}^1$. If the frames are well classified then the noise frames only include negligible speech power, due to the sparsity and non-stationarity of speech; the speech frames include noise power similar to the noise frames, due to the stationarity of noise. Therefore, inter-frame spectral subtraction can be performed as follows: for each speech frame, the cross- and auto-PSD of its nearest noise frame is subtracted from its cross- and auto-PSD, then its noise-free cross- and auto-PSD is obtained and denoted as $\hat{\boldsymbol{\phi}}_{t,f}^i$.

Instead of using $\mathbf{x}_{t,f}^i$, we use $\hat{\boldsymbol{\phi}}_{t,f}^i$ to construct (3). Correspondingly, we have a new formula (4), which is still valid, since it is equivalent to, with noise removed, taking the cross- and auto-PSD between both sides of the initial formula (4) and $x_{t,f}^1$. In the RLS process, only the speech frames (after spectral

subtraction) are used, and the noise frames are skipped. A speech frame with a preceding noise frame is initialized with the latest speech frame.

3) Consistency test: In practice, a DP-RTF estimate can sometimes be unreliable. Possible reasons are that in a small frame region, (i) the CTF is time-varying due to a fast movement of the speakers, (ii) multiple speakers are present, (iii) only noise is present due to a wrong noise-speech classification, or (iv) only reverberation is present at the end of a speech occurrence. In [12], a consistency test was proposed to tackle this problem: If a small frame region indeed corresponds to one active speaker, the DP-RTFs estimated using different reference channels are consistent, otherwise the DP-RTFs are biased, with inconsistent bias values. In the present work, we use the first and second channels as references, we obtain the DP-RTF estimates $\tilde{c}^i_{t,f}$ (with $\tilde{c}^1_{t,f}=1$) and $\bar{c}^i_{t,f}$ (with $\bar{c}^2_{t,f}=1$), respectively. Then $\tilde{c}^i_{t,f}$ and $\bar{c}^i_{t,f}/\bar{c}^1_{t,f}$ are two estimates of the same DP-RTF $a^i_{0,f}/a^1_{0,f}$. To measure the similarity between these two estimates, we define the vectors $\mathbf{c}^i_{1,t,f}=(1,\tilde{c}^i_{t,f})^\top$ and $\mathbf{c}^i_{2,t,f}=(1,\bar{c}^i_{t,f}/\bar{c}^1_{t,f})^\top$, where the first entries are the DP-RTFs corresponding to $a^1_{0,f}/a^1_{0,f}=1$. The similarity is the cosine of the angle between the two unit vectors:

$$d^i_{t,f}=\frac{|\mathbf{c}^{i\ H}_{1,t,f}\mathbf{c}^i_{2,t,f}|}{\sqrt{\mathbf{c}^{i\ H}_{1,t,f}\mathbf{c}^i_{1,t,f}\mathbf{c}^{i\ H}_{2,t,f}\mathbf{c}^i_{2,t,f}}}, \tag{11}$$

where ^{H} denotes conjugate transpose. If $d^i_{t,f}\in[0,1]$ is larger than a threshold (which is fixed to 0.75 in this work) then the two estimates are consistent, otherwise they are simply ignored. Then, the two estimates are averaged and normalized as done in [12], resulting in a final complex-valued feature $\hat{c}^i_{t,f}$ whose module lies in the interval $[0,1]$.

Finally, at frame t, we obtain a set of features $\mathcal{C}_t=\{\{\hat{c}^i_{t,f}\}_{i\in\mathcal{I}_f}\}_{f=0}^{F-1}$, where $\mathcal{I}_f\subseteq\{2,\dots,I\}$ denotes the set of microphone indices that pass the consistency test. Note that $\mathcal{I}_f$ is empty if frame t is a noise frame at frequency f, or if no channel passes the consistency test. Each one of these features is assumed to be associated with only one speaker.

III. Localization of Multiple Moving Speakers

In this section we describe the proposed frame-wise online multiple-speaker localizer. We start by briefly presenting the underlying complex Gaussian mixture model, followed by the recursive estimation of its parameters.

A. Generative Model for Multiple-Speaker Localization

In order to associate DP-RTF features from $\mathcal{C}_t$ with speakers and to localize each active speaker, we adopt the generative model proposed in [7]. Let $\mathcal{D}=\{\tilde{\theta}_1,\dots,\tilde{\theta}_d,\dots,\tilde{\theta}_D\}$ be a set of D candidate source *directions*, e.g. azimuth angles. An observed feature $\hat{c}^i_{t,f}$ (cf. Section II), when emitted by a sound source located along the direction $\tilde{\theta}_d$, is assumed to be drawn from a complex-Gaussian distribution with mean $c^{i,d}_f$ and variance σ^2, i.e. $\hat{c}^i_{t,f}|d\sim\mathcal{N}_c(c^{i,d}_f,\sigma^2)$. The mean $c^{i,d}_f$ is the predicted feature at frequency f for channel i, and is precomputed based on direct-path propagation along azimuth $\tilde{\theta}_d$ to the microphones. The variance σ^2 is empirically set as a constant value. The marginal density of an observed feature $\hat{c}^i_{t,f}$ (taking into account all candidate directions) is a CGMM with each component corresponding to a candidate direction:

$$p(\hat{c}^i_{t,f}|\mathcal{D})=\sum_{d=1}^{D}w_d\mathcal{N}_c(\hat{c}^i_{t,f};c^{i,d}_f,\sigma^2), \tag{12}$$

where $w_d\geq 0$ is the prior probability (component weight) of the d-th component, with $\sum_{d=1}^{D}w_d=1$. Let us denote the vector of weights with $\boldsymbol{w}=(w_1,\dots,w_D)^\top$. Note that this vector is the only free parameter of the model.

Assuming that the observations in $\mathcal{C}_t$ are independent, the corresponding (normalized) negative log-likelihood function, as a function of w_d, is given by:

$$\mathcal{L}_t=-\frac{1}{|\mathcal{C}_t|}\sum_{\hat{c}^i_{t,f}\in\mathcal{C}_t}\log\Big(\sum_{d=1}^{D}w_d\mathcal{N}_c(\hat{c}^i_{t,f};c^{i,d}_f,\sigma^2)\Big), \tag{13}$$

where $|\mathcal{C}_t|$ denotes the cardinality of $\mathcal{C}_t$. Once $\mathcal{L}_t$ is minimized, each weight w_d represents the probability that a speaker is active in the direction $\tilde{\theta}_d$. Therefore, sound source localization amounts to the minimization of $\mathcal{L}_t$. In addition, taking into account the fact that the number of actual active speakers is much lower than the number of candidate directions, an entropy term was proposed in [12] as a regularizer to impose a sparse solution for w_d. The entropy is defined as

$$H=-\sum_{d=1}^{D}w_d\log(w_d). \tag{14}$$

The concave-convex procedure [34] was adopted in [12], to minimize the objective function $\mathcal{L}+\gamma H$ w.r.t. $\boldsymbol{w}$, where $\mathcal{L}$ is the normalized negative log-likelihood of the DP-RTF features of all frames, i.e. batch mode optimization, and the positive scalar γ was used to control the tradeoff between likelihood minimization and imposing sparsity over the weights. In the batch mode, the weight vector $\boldsymbol{w}$ is shared across all frames. Hence this method is not suitable for moving speakers.

B. Recursive Parameter Estimation

We now describe a recursive method for updating the weight vector from $\boldsymbol{w}_{t-1}$ to $\boldsymbol{w}_t$, i.e. from frame $t-1$ to frame t, using the DP-RTF features at t. This can be formulated as the following online optimization problem [31]:

$$\boldsymbol{w}_t=\underset{\boldsymbol{w}}{\operatorname{argmin}}\ \chi(\boldsymbol{w},\boldsymbol{w}_{t-1})+\eta(\mathcal{L}_t+\gamma H), \tag{15}$$

$$\text{s.t.}\quad w_d>0,\ \forall d\in\{1\dots D\}\quad\text{and}\quad\sum_{d=1}^{D}w_d=1, \tag{16}$$

where $\chi(\boldsymbol{a},\boldsymbol{b})$ is a distance between $\boldsymbol{a}$ and $\boldsymbol{b}$. The positive scalar factor η controls the parameter update rate. To minimize (15), the derivative of the objective function w.r.t $\boldsymbol{w}$ is set to zero, yielding a set of equations with no closed-form solution. To speed up the computation, it is assumed that $\boldsymbol{w}_t$

is close to $\boldsymbol{w}_{t-1}$, thence the derivative of $\mathcal{L}_t + \gamma H$ at $\boldsymbol{w}$ can be approximated with the derivative of $\mathcal{L}_t + \gamma H$ at $\boldsymbol{w}_{t-1}$. This assumption is reasonable when parameter evolution is not too fast. As a result, when the distance $\chi(\boldsymbol{w}, \boldsymbol{w}_{t-1})$ is Euclidean, the objective function leads to gradient descent with a step length equal to η. Nevertheless, the constraints (16) lead to an inefficient gradient descent procedure. To obtain an efficient solver, we exploit the fact that the weights w_d are probability masses, hence we replace the Euclidean distance with the more suitable Kullback-Leibler divergence, i.e. $\chi(\boldsymbol{w}, \boldsymbol{w}_{t-1}) = \sum_{d=1}^{D} w_d \log \frac{w_d}{w_{t-1,d}}$, which results in the exponentiated gradient algorithm [31].

The partial derivatives of $\mathcal{L}_t$ and H w.r.t w_d at the point $w_{t-1,d}$ are computed with, respectively:

$$\begin{aligned} \frac{\partial(\mathcal{L}_t)}{\partial w_d}\Big|_{w_{t-1,d}} &= -\frac{1}{|\mathcal{C}_t|} \sum_{\hat{c}^i_{t,f} \in \mathcal{C}_t} \frac{\mathcal{N}_c(\hat{c}^i_{t,f}; c_f^{i,d}, \sigma^2)}{\sum_{d'=1}^{D} w_{t-1,d'} \mathcal{N}_c(\hat{c}^i_{t,f}; c_f^{i,d'}, \sigma^2)}, \\ \frac{\partial H}{\partial w_d}\Big|_{w_{t-1,d}} &= -(1 + \log(w_{t-1,d})), \quad \forall d \in \{1 \dots D\}. \end{aligned} \tag{17}$$

Then, the exponentiated gradient,

$$r_{t-1,d} = e^{-\eta\left(\frac{\partial(-\mathcal{L}_t)}{\partial w_d}\Big|_{w_{t-1,d}} + \gamma \frac{\partial H}{\partial w_d}\Big|_{w_{t-1,d}}\right)}, \quad \forall d \in \{1 \dots D\}, \tag{18}$$

is used to update the weights with:

$$w_{t,d} = \frac{r_{t-1,d} w_{t-1,d}}{\sum_{d'=1}^{D} r_{t-1,d'} w_{t-1,d'}}, \quad \forall d \in \{1 \dots D\}. \tag{19}$$

It is clear from (19) that the parameter constraints (16) are necessarily satisfied. The exponentiated gradient algorithm sequentially evaluates (17), (18) and (19) at each frame. At the first frame, the weights are initialized with the uniform distribution, namely $w_{1,d} = \frac{1}{D}$. When $\mathcal{C}_t$ is empty, such as during a silent period, the parameters are recursively updated with $w_{t,d} = (1 - \eta')w_{t-1,d} + \eta' \frac{1}{D}$.

The weight $\boldsymbol{w}_t$ as a function of $\tilde{\theta}_d$, i.e. $w_{t,d}$, exhibits a handful of peaks that could correspond to active speakers. The use of an entropy regularization term was shown to both suppress small spurious peaks, present without using the regularization term, and to sharpen the peaks corresponding to actual active speakers, thus allowing to better localize true speakers and to eliminate erroneous ones. In the case of moving speakers, a peak should shift along time from a direction $\tilde{\theta}_d$ to a nearby direction. Spatial smoothing of the weight function raises the weight values around a peak, which results in smoother peak jumps. In our experiments, spatial smoothing is carried out with $w_{t,d} = (w_{t,d} + 0.02 w_{t,d-1} + 0.02 w_{t,d+1})/1.04$, where the smoothing factor 0.02 is empirically chosen in order to smooth peak motion from one frame to the next, while avoiding the peaks to collapse. One may think that spatial smoothing and entropy regularization neutralize each other, but in practice it was found that their combination is beneficial.

C. Peak Selection and Frame-wise Speaker Localization

Frame-wise localization and counting of active speakers could be carried out by selecting the peaks of $\boldsymbol{w}_t(\tilde{\theta}_d)$ larger than a predefined threshold [12], [30]. However, peak selection does not exploit the temporal dependencies of moving speakers. Moreover, peak selection can be a risky process since a too high or too low threshold value may lead to undesirable missed detection or false alarm rates. In order to avoid these problems, we adopt a weighted-data Bayesian framework: all the candidate directions and the associated weights are used as observations by the multiple speaker tracking method described in Section IV below. The localization results obtained with peak selection are compared with the localization results obtained with the proposed tracker in Section V.

IV. Multiple Speaker Tracking

Let N be the maximum number of speakers that can be simultaneously active at any time t, and let n be the speaker index. Moreover, let $n = 0$ denote *no speaker*. We now introduce the main variables and their notations. Upper case letters denote random variables while lower case letters denote their realizations.

Let $\boldsymbol{S}_{tn}$ be a latent (or state) variable associated with speaker n at frame t, and let $\mathbf{S}_t = (\boldsymbol{S}_{t1}, \dots, \boldsymbol{S}_{tn}, \dots, \boldsymbol{S}_{tN})$. $\boldsymbol{S}_{tn}$ is composed of two parts: the speaker direction and the speaker velocity. In this work, speaker direction is defined by an azimuth θ_{tn}. To avoid phase (circular) ambiguity we describe the direction with the unit vector $\boldsymbol{U}_{tn} = (\cos(\theta_{tn}), \sin(\theta_{tn}))^\top$. Moreover, let $V_{tn} \in \mathbb{R}$ be the angular velocity. Altogether we define a realization of the state variable as $\boldsymbol{s}_{tn} = [\boldsymbol{u}_{tn}; v_{tn}]$ where the notation $[\cdot;\cdot]$ stands for vertical vector concatenation.

Let $\mathbf{O}_t = (\boldsymbol{O}_{t1}, \dots, \boldsymbol{O}_{td}, \dots, \boldsymbol{O}_{tD})$ be the observed variables at frame t. Each realization $\boldsymbol{o}_{td}$ of $\boldsymbol{O}_{td}$ is composed of a candidate location, or azimuth $\tilde{\theta}_{td} \in \mathcal{D}$, and a weight w_{td}. The weight w_{td} is the probability that there is an active speaker in the direction $\tilde{\theta}_{td}$, namely (15). As above, let the azimuth be described by a unit vector $\boldsymbol{b}_{td} = (\cos(\tilde{\theta}_{td}), \sin(\tilde{\theta}_{td}))^\top$. In summary we have $\boldsymbol{o}_{td} = [\boldsymbol{b}_{td}; w_{td}]$. Moreover, let Z_{td} be a (latent) assignment variable associated with each observed variable $\boldsymbol{O}_{td}$, such that $Z_{td} = n$ means that the observation indexed by d at frame t is assigned to active speaker $n \in \{0, \dots, N\}$. Note that $Z_{td} = 0$ is a "fake" assignment – the corresponding observation is assigned to an audio source that is either background noise or any other source that has not yet been identified as an active speaker.

The problem at hand can now be cast into the estimation of the filtering distribution $p(\mathbf{s}_t, \boldsymbol{z}_t | \mathbf{o}_{1:t})$, and further inference of $\mathbf{s}_t$ and $\boldsymbol{z}_t$. In this work we make two hypotheses, namely (i) that the observations at frame t only depend on the assignment and state variables at t, and (ii) that the prior distribution of the assignment variables is independent of all the other variables. By applying the Bayes rule together with these hypotheses,

and ignoring terms that do not depend on $\mathbf{s}_t$ and $\boldsymbol{z}_t$, the filtering distribution is proportional to:

$$p(\mathbf{s}_t, \boldsymbol{z}_t|\mathbf{o}_{1:t}) \propto p(\mathbf{o}_t|\mathbf{s}_t, \boldsymbol{z}_t)p(\boldsymbol{z}_t)p(\mathbf{s}_t|\mathbf{o}_{1:t-1}), \quad (20)$$

which contains three terms: the observation model, the prior distribution of the assignment variable and the predictive distribution over the sources state. We now characterize each one of these three terms.

1) Audio observation model: The audio observation model describes the distribution of the observations given speakers state and assignment. We assume the different observations are independent, conditioned on speakers state and assignment, which can be written as:

$$p(\mathbf{o}_t|\boldsymbol{z}_t, \mathbf{s}_t) = \prod_{d=1}^{D} p(\mathbf{o}_{td}|\boldsymbol{z}_t, \mathbf{s}_t). \quad (21)$$

Since the weights describe the confidence associated with each observed azimuth, we adopt the weighted-data GMM model of [35]:

$$p(\boldsymbol{b}_{td}|Z_{td} = n, \mathbf{s}_{tn}; w_{td}) = \begin{cases} \mathcal{N}(\boldsymbol{b}_{td}; \mathbf{M}\boldsymbol{s}_{tn}, \frac{1}{w_{td}}\boldsymbol{\Sigma}) & \text{if } n \in \{1, \dots, N\} \\ \mathcal{U}(\text{vol}(\mathcal{G})) & \text{if } n = 0 \end{cases}, \quad (22)$$

where the matrix $\mathbf{M} = [\mathbf{I}_{2\times 2}, \mathbf{0}_{2\times 1}]$ projects the state variable onto the space of source directions and $\boldsymbol{\Sigma}$ is a covariance matrix (set empirically to a fixed value in the present study). Note that the weight plays the role of a precision: The higher the weight w_{td}, the more reliable the source direction $\boldsymbol{b}_{td}$. The case $Z_{td} = 0$ follows a uniform distribution over the volume of the observation space.

2) Prior distribution: The prior distribution of the assignment variable is independent over observations and is assumed to be uniformly distributed over all the speakers (including the case $n = 0$), hence:

$$p(\boldsymbol{z}_t) = \prod_{d=1}^{D} p(Z_{td} = n) \quad \text{with} \quad \pi_{dn} = p(Z_{td} = n) = \frac{1}{N+1}. \quad (23)$$

3) Predictive distribution: The predictive distribution describes the relationship between the state $\mathbf{s}_t$ and the past observations up to frame t, $\mathbf{o}_{1:t-1}$. To calculate this distribution, we first marginalize $p(\mathbf{s}_t|\mathbf{o}_{1:t-1})$ over $\mathbf{s}_{t-1}$, writing:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) = \int p(\mathbf{s}_t|\mathbf{s}_{t-1})p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})d\mathbf{s}_{t-1}, \quad (24)$$

where the two terms under the integral are the state dynamics and the marginal filtering distribution of the state variable at frame $t-1$, respectively. We model the state dynamics as a linear-Gaussian first-order Markov process, independent over the speakers, i.e. :

$$p(\mathbf{s}_t|\mathbf{s}_{t-1}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}_{t-1,n}\boldsymbol{s}_{t-1,n}, \boldsymbol{\Lambda}_{tn}), \quad (25)$$

where Λ_{tn} is the dynamics' covariance matrix and $\mathbf{D}_{t-1,n}$ is the state transition matrix. Given the estimated azimuth $\theta_{t-1,n}$ and angular velocity $v_{t-1,n}$ at frame $t-1$, we have the following relation:

$$\begin{pmatrix} \cos(\theta_{tn}) \\ \sin(\theta_{tn}) \end{pmatrix} = \begin{pmatrix} \cos(\theta_{t-1,n} + v_{t-1,n}\Delta t) \\ \sin(\theta_{t-1,n} + v_{t-1,n}\Delta t) \end{pmatrix}, \quad (26)$$

where Δt is the time increment between two consecutive frames. Expanding (26) and assuming that the angular displacement $v_{t-1,n}\Delta t$ is small, the state transition matrix can be written as:

$$\mathbf{D}_{t-1,n} = \begin{pmatrix} 1 & 0 & -\sin(\theta_{t-1,n})\Delta t \\ 0 & 1 & \cos(\theta_{t-1,n})\Delta t \\ 0 & 0 & 1 \end{pmatrix}. \quad (27)$$

In the following $\mathbf{D}_{t-1,n}$ is written as $\mathbf{D}$, only to lighten the equations.

A. Variational Expectation Maximization Algorithm

At this point, the standard solution to the calculation of the filtering distribution consists of using EM methodology. EM alternates between evaluating the expected complete-data log-likelihood and maximizing this expectation with respect to the model parameters. More precisely, the expectation writes:

$$J(\boldsymbol{\Theta}, \boldsymbol{\Theta}^o) = \mathbf{E}_{p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t}, \boldsymbol{\Theta}^o)} \left[\log p(\boldsymbol{z}_t, \mathbf{s}_t, \mathbf{o}_{1:t}|\boldsymbol{\Theta})\right], \quad (28)$$

where $\boldsymbol{\Theta}^o$ denotes the current parameter estimates and $\boldsymbol{\Theta}$ denotes the new estimates, obtained via maximization of (28). However, given the hybrid combinatorial-and-continuous nature of the latent space, such solution is intractable in practice, due to combinatorial explosion. We thus propose to use of a variational approximation to solve the problem efficiently. We inspire from [22] and propose the following factorization:

$$p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t}) \approx q(\boldsymbol{z}_t, \mathbf{s}_t) = q(\boldsymbol{z}_t) \prod_{n=0}^{N} q(\boldsymbol{s}_{tn}). \quad (29)$$

The optimal solution is then given by two E-steps, an E-S step for each individual state variable $\boldsymbol{S}_{tn}$ and an E-Z step for the assignment variable $\boldsymbol{Z}_t$:

$$\log q(\boldsymbol{s}_{tn}) = \mathbf{E}_{q(\boldsymbol{z}_t)\prod_{m\neq n} q(\boldsymbol{s}_{tm})}[\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})], \quad (30)$$

$$\log q(\boldsymbol{z}_t) = \mathbf{E}_{q(\mathbf{s}_t)}[\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})]. \quad (31)$$

It is easy to see that in order to compute (30) and (31), two elements are needed: the predictive distribution (24) and the marginal filtering distribution at $t-1$, $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$. Remarkably, as a consequence of the factorization (29), we can replace $p(\mathbf{s}_{t-1}|\mathbf{o}_{1:t-1})$ with $q(\mathbf{s}_{t-1}) = \prod_{n=1}^{N} q(\boldsymbol{s}_{t-1,n})$ in (24) and compute the predictive distribution as follows:

$$p(\mathbf{s}_t|\mathbf{o}_{1:t-1}) \approx \int p(\mathbf{s}_t|\mathbf{s}_{t-1}) \prod_{n=1}^{N} q(\boldsymbol{s}_{t-1,n})d\mathbf{s}_{t-1}. \quad (32)$$

This predictive distribution factorizes across speakers. Moreover, one prominent feature of the proposed variational approximation is that, if the posterior distribution at time $t-1$ $q(\boldsymbol{s}_{t-1,n})$ is assumed to be a Gaussian, say

$$q(\boldsymbol{s}_{t-1,n}) = \mathcal{N}(\boldsymbol{s}_{t-1,n}; \boldsymbol{\mu}_{t-1,n}, \boldsymbol{\Gamma}_{t-1,n}), \quad (33)$$

then (the approximation of) the predictive distribution (32) is a Gaussian. More specifically, the derivation of (32) leads to:

$$p(\mathbf{s}_{tn}|\mathbf{o}_{1:t-1}) = \mathcal{N}(\mathbf{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}). \quad (34)$$

Moreover, as we will see in the E-S-step below, the posterior distribution at time t, $q(\boldsymbol{s}_{tn})$, is also a Gaussian.

1) E-S step: The computation of the variational posterior distribution $q(\boldsymbol{s}_{tn})$, for all currently tracked speakers, is carried out by developing (30) as follows. We first exploit (20), (21), (23) and (34) to rewrite $\log p(\boldsymbol{z}_t, \mathbf{s}_t|\mathbf{o}_{1:t})$ in (30) as a sum of individual log-probabilities. Then we eliminate all terms not depending on $\boldsymbol{s}_{tn}$ and we evaluate the expectation of the remaining terms. Because the terms not depending on $\boldsymbol{s}_{tn}$ were disregarded, the expectation is computed only with respect to $q(\boldsymbol{z}_t)$. This nicely makes the computation of $q(\boldsymbol{s}_{tn})$ independent of the structure of $q(\boldsymbol{s}_{tm})$ for $m \neq n$. Eventually, this yields a Gaussian distribution:

$$q(\boldsymbol{s}_{tn}) = \mathcal{N}(\boldsymbol{s}_{tn}; \boldsymbol{\mu}_{tn}, \boldsymbol{\Gamma}_{tn}), \quad (35)$$

with the following parameters:

$$\boldsymbol{\Gamma}_{tn} = \Big(\Big(\sum_{d=1}^{D} \alpha_{tdn} w_{td}\Big)\mathbf{M}^\top \boldsymbol{\Sigma}^{-1}\mathbf{M} + \Big(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top\Big)^{-1}\Big)^{-1}, \quad (36)$$

$$\boldsymbol{\mu}_{tn} = \boldsymbol{\Gamma}_{tn}\Big(\mathbf{M}^\top\boldsymbol{\Sigma}^{-1}\Big(\sum_{d=1}^{D}\alpha_{tdn}w_{td}\boldsymbol{b}_{td}\Big) + \Big(\boldsymbol{\Lambda}_{tn} + \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top\Big)^{-1}\mathbf{D}\boldsymbol{\mu}_{t-1,n}\Big), \quad (37)$$

where $\alpha_{tdn} = q(Z_{td} = n)$ is the variational posterior distribution of the assignment variable, which will be detailed in Section IV-A2. Importantly, the first two entries of $\boldsymbol{\mu}_{tn}$ in (37) represent the estimated azimuth of speaker n. The "final" azimuth estimate at frame t is thus given by this subvector at the end of the VEM iterations. Since we use a unit-vector representation, we normalize this vector at each iteration of the algorithm. Finally, note that since we have shown that $q(\boldsymbol{s}_{t-1,n})$ being Gaussian leads to $q(\boldsymbol{s}_{tn})$ being Gaussian as well, it is sufficient to assume that $q(\boldsymbol{s}_{1n})$ is Gaussian, namely at $t = 1$: $q(\boldsymbol{s}_{1n}) = \mathcal{N}(\boldsymbol{s}_{1n}; \boldsymbol{\mu}_{1n}, \boldsymbol{\Gamma}_{1n})$.

2) E-Z step: Developing (31) with the same principles as above, one can easily find that the variational posterior distribution of the assignment variable factorizes as:

$$q(\boldsymbol{z}_t) = \prod_{d=1}^{D} q(z_{td}). \quad (38)$$

In addition, we obtain a closed-form expression for $q(z_{td})$:

$$\alpha_{tdn} = q(Z_{td} = n) = \frac{\rho_{tdn}\pi_{dn}}{\sum_{i=0}^{N}\rho_{tdi}\pi_{di}}, \quad (39)$$

where π_{dn} was defined in (23), and ρ_{tdn} is given by:

$$\rho_{tdn} = \begin{cases} \mathcal{N}(\boldsymbol{b}_{td}; \mathbf{M}\boldsymbol{\mu}_{tn}, \frac{1}{w_{td}}\boldsymbol{\Sigma}) \\ \times e^{-\frac{1}{2}\mathrm{tr}\left(w_{td}\mathbf{M}^\top\boldsymbol{\Sigma}^{-1}\mathbf{M}\boldsymbol{\Gamma}_{tn}\right)} & \text{if } 1 \leq n \leq N \\ \mathcal{U}(\mathrm{vol}(\mathcal{G})) & \text{if } n = 0. \end{cases} \quad (40)$$

3) M-step: Once the two expectation steps are executed, we maximize J in (28) with respect to the model parameters, i.e. the covariance matrix of the state dynamics $\boldsymbol{\Lambda}_{tn}$. By exploiting again the proposed variational approximation, the dependency of J on $\boldsymbol{\Lambda}_{tn}$ can be written as:

$$J(\boldsymbol{\Lambda}_{tn}) = \mathbf{E}_{q(\boldsymbol{s}_{tn})}\left[\log\mathcal{N}(\boldsymbol{s}_{tn}; \mathbf{D}\boldsymbol{\mu}_{t-1,n}, \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn})\right],$$

which can be further developed as:

$$J(\boldsymbol{\Lambda}_{tn}) = \log|\mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn}| + \mathrm{Tr}\left[(\mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + \boldsymbol{\Lambda}_{tn})^{-1} \times ((\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1,n})^\top + \boldsymbol{\Gamma}_{tn})\right]. \quad (41)$$

By equating to zero the gradient of (41) w.r.t. $\boldsymbol{\Lambda}_{tn}$, we obtain:

$$\boldsymbol{\Lambda}_{tn} = \boldsymbol{\Gamma}_{tn} - \mathbf{D}\boldsymbol{\Gamma}_{t-1,n}\mathbf{D}^\top + (\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1,n})(\boldsymbol{\mu}_{tn} - \mathbf{D}\boldsymbol{\mu}_{t-1,n})^\top. \quad (42)$$

B. Speaker-Birth Process

A birth process is used to initialize new tracks, i.e. speakers that become active. We take inspiration from the birth process for visual tracking proposed in [22] and adapt it to audio tracking. The general principle is the following. In a short period of time, say from $t - L$ to t, with L being small (typically 3), we assume that at most one new (yet untracked) speaker becomes active. For each frame from $t - L$ to t, among the observations assigned to $n = 0$ we select the one with the highest weight, and thus obtain an observation sequence $\tilde{\boldsymbol{o}}_{t-L:t}$. We then compute the marginal likelihood of this sequence according to our model, $\epsilon_0 = p(\tilde{\boldsymbol{o}}_{t-L:t})$. If these observations have been generated by a speaker that has not been detected yet (hypothese H_1), then they are assumed to be consistent with the model, i.e. exhibit smooth trajectories, and ϵ_0 will be high; otherwise, i.e. if they have been generated by background noise (hypothese H_0), they will be more randomly spread over the range of possible observations, and ϵ_0 will be low. Giving birth to a new speaker track amounts to setting a threshold ϵ_1 and deciding between the two hypotheses:

$$\epsilon_0 \underset{H_0}{\overset{H_1}{\gtrless}} \epsilon_1. \quad (43)$$

This process is applied continuously over time to detect new speakers. This includes speaker track initialization at $t = 1$. Note that initially all the assignment variables are set to $n = 0$ (background noise), namely $Z_{1d} = 0, \forall d$.

As for the computaiotn of $p(\tilde{\boldsymbol{o}}_{t-L:t})$, we first rewrite it as the marginalization of the joint probability of the selected observations and the state trajectory $\hat{\boldsymbol{s}}_{t-L:t}$ of a potential speaker:

$$\epsilon_0 = \int p(\tilde{\boldsymbol{o}}_{t-L:t}, \hat{\boldsymbol{s}}_{t-L:t})d\hat{\boldsymbol{s}}_{t-L:t}, \quad (44)$$

which, under the proposed model, is given by:

$$\epsilon_0 = \int\Big(\prod_{i=t-L+1}^{t} p(\tilde{\boldsymbol{o}}_i|\hat{\boldsymbol{s}}_i)p(\hat{\boldsymbol{s}}_i|\hat{\boldsymbol{s}}_{i-1})\Big)p(\tilde{\boldsymbol{o}}_{t-L}|\hat{\boldsymbol{s}}_{t-L})p(\hat{\boldsymbol{s}}_{t-L})d\hat{\boldsymbol{s}}_{t-L:t}. \quad (45)$$

Algorithm 2 Variational EM tracking

```
Input: audio observations b_{1:t}
for t = 1 to end do
  Gather observations at frame t
  for iter = 1 to N_iter do
    E-Z-step:
    for d ∈ {1, ..., D} do
      for n ∈ {0, ..., N} do
        Evaluate q(Z_td = n) with (39)
      end for
    end for
    E-S-step:
    for n ∈ {1, ..., N} do
      Evaluate Γ_tn and μ_tn with (36) and (37);
    end for
    M-step: Evaluate Λ_tn with (42);
  end for
  Speaker-Birth Process (see Section IV-B)
  Detect speaker activity (see Section IV-C)
  for n ∈ {1, ..., N} do
    if the speaker n is detected as active then
      Output the results μ_tn
    end if
  end for
end for
```

All the terms in the above equation have been defined during the derivation of our model except the marginal prior distribution of the state $p(\hat{\boldsymbol{s}}_{t-L})$, and all these terms are Gaussian. For the track-birth process, we just want to test if the trajectory of observations from $t-L$ to t is coherent, and we can define here $p(\hat{\boldsymbol{s}}_{t-L})$ as a non-informative distribution, such as a uniform distribution. In practice we choose a Gaussian distribution with a very large covariance, to ensure a closed-form solution to (45). Due to room limitation, we do not present more details. Let us just mention that in practice we set $L = 3$, which enables efficient speaker birth detection.

C. Speaker Activity Detection

A very interesting feature of the proposed model is that, once speaker tracks have been estimated, the posterior distribution of the assignment variables $\boldsymbol{Z}_t$ can be used for speech activity detection, i.e. who are the active speakers at each frame, a task also referred to as *speaker diarization* in the multi-speaker context. This can be formalized as testing for each frame t and each speaker n between the two following hypotheses: H_1: Speaker n is active at frame t, and H_0: Speaker n is silent at frame t. In the present work, this is done by computing the following *weighted sum of weights*, averaged over a small number of frames L' to take into account speaker activity inertia, and comparing with a threshold δ, a test formally written as:

$$\sum_{i=t-L'+1}^{t} \sum_{d=1}^{D} \alpha_{idn} w_i^d \underset{H_0}{\overset{H_1}{\gtrless}} \delta. \tag{46}$$

Overall, the variational EM tracking algorithm is described in Algorithm 2.

V. Experiments

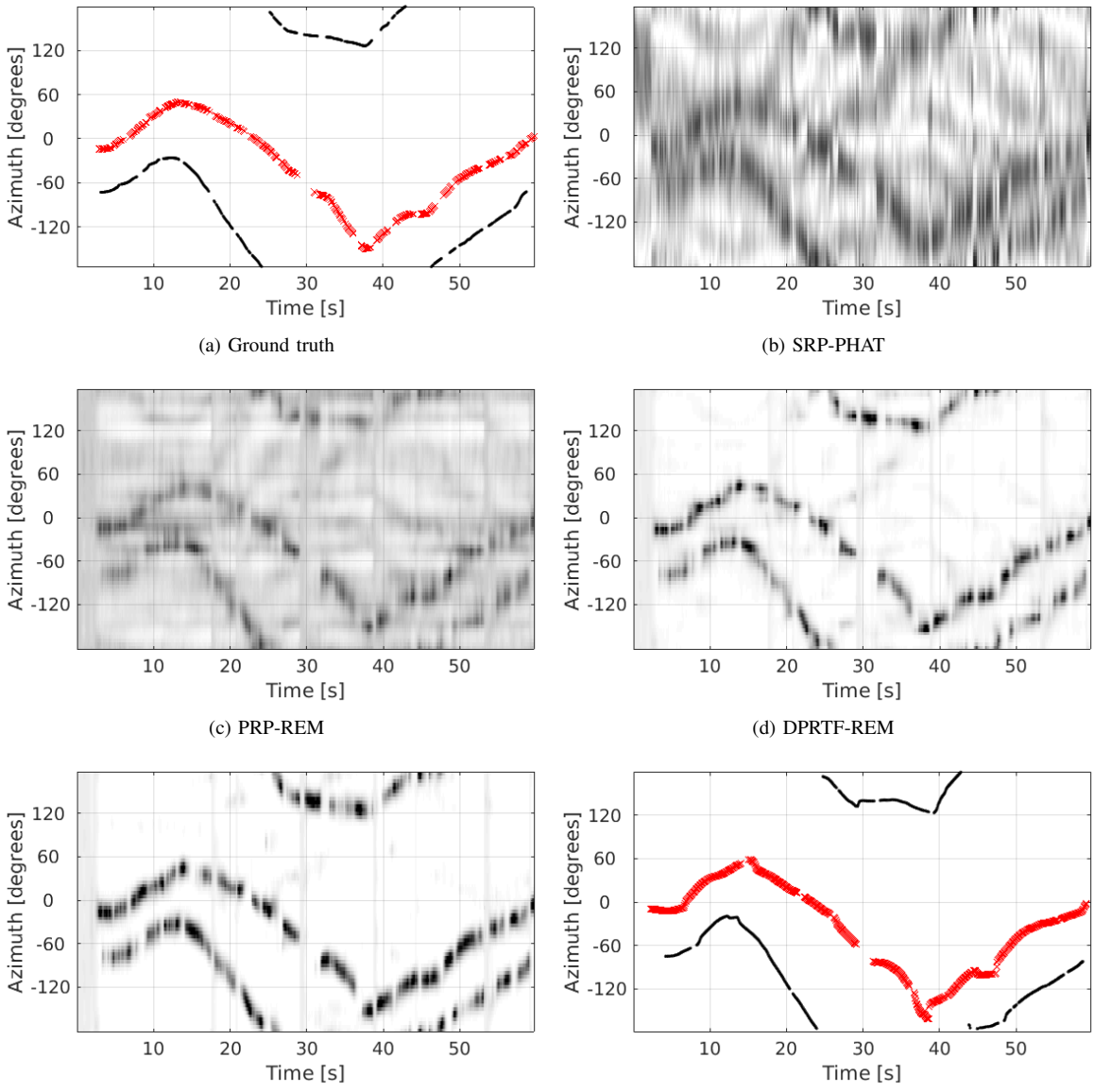
A. Experimental setups

1) Datasets: We tested and empirically validated our method with the LOCATA and the Kinovis multiple speaker tracking (Kinovis-MST) datasets. The LOCATA (a IEEE-AASP challenge for sound source localization and tracking) [36] data were recorded in the Computing Laboratory of the Department of Computer Science of Humboldt University Berlin. The room size is 7.1 m $\times$ 9.8 m $\times$ 3 m, with a reverberation time $T_{60} \approx 0.55$ s. We report the results of the development corpus for tasks #3 and #5 with a single moving speaker, and for tasks #4 and #6 with two moving speakers, each task comprising three recorded sequences.[2] There are twelve microphones arranged such as to form a spherical array and placed on the head of a NAO robot. We used two microphone configurations: four quasi-planar microphones, located on the top of the head, numbered 5, 8, 11, 12, and eight microphones numbered 1, 3, 4, 5, 8, 10, 11, 12. An optical motion capture system was used to provide ground-truth positions of the robot and of the speakers. The participants speak continuously during the entire recordings. However, speech pauses are inevitable and these pauses may last several seconds. Each participant has a head-mounted microphone. We applied the voice activity detector [37] to these microphone signals to obtain ground-truth voice activity information of each participant. The signal-to-noise ratio (SNR) is approximatively 23.4 dB

The Kinovis-MST dataset was recorded in the Kinovis multiple-camera laboratory at INRIA Grenoble.[3] The room size is 10.19 m $\times$ 9.87 m $\times$ 5.6 m, with $T_{60} \approx 0.53$ s. A v5 NAO robot with four microphones [38] was used. The geometric layout of the microphones is similar to the one of the robot used in LOCATA. The speakers were moving around the robot with a speaker-to-robot distance ranging between 1.5 m and 3.5 m. As with LOCATA, a motion capture system was used to obtain ground-truth trajectories of the moving participants and the location of the robot. Ten sequences were recorded with up to three participants, for a total length of about 357 s. The robot's head has built-in fans located nearby the microphones, hence the recordings contain a notable amount of stationary and spatially correlated noise with an SNR of approximatively 2.7 dB [38]. The participants behave more naturally than in the LOCATA scenarios, i.e. they take speech turns in a natural multi-party dialog. When one participant is silent, he/she manually hides the infrared marker located on his head to make it invisible to the motion capture system. This provides ground-truth speech activity information for each

[2]The results obtained with the proposed method were officially submitted to the LOCATA challenge and they will be available soon at https://locata.lms.tf.fau.de/.

[3]The Kinovis-MST dataset is publicly available at: https://team.inria.fr/perception/the-kinovis-mst-dataset/



(a) Ground truth

(b) SRP-PHAT

(c) PRP-REM

(d) DPRTF-REM

(e) DPRTF-EG

(f) VEM-tracking

Fig. 2: Results of speaker localization and tracking for Recording 1 / Task 6 of LOCATA data. (a) Ground truth trajectory and voice activity (red for speaker 1, black for speaker 2). Intervals in the trajectories are speaking pauses. (b)-(e) One-dimensional heat maps as a function of time for the four tested localization methods. (f) Results for the proposed VEM-based tracker. Black and red colors demonstrate a successful tracking, i.e. continuity of the tracks despite of speech pauses.

participant. This dataset and the associated annotations allow us to test the proposed tracking algorithm when the number of active speakers varies over time.

2) Parameter setting: For both datasets, we perform 360°-wide azimuth estimation and tracking: $D = 72$ azimuth directions at every 5° in $[-175°, 180°]$ are used as candidate directions. The CGMM mean $c_f^{i,d}$ is the head-related transfer function (HRTF) ratio between two microphones, which are precomputed based on the direct-path propagation model for each candidate direction. In the Kinovis-MST dataset, the HRTFs have been measured to compute the CGMM means. For LOCATA, the TDOAs are computed based on the coordinate of microphones, which are then used to compute the phase of the CGMM means, while the magnitude of the CGMM means are set to a constant, e.g. 0.5, for all the frequencies. All the recorded signals are resampled to 16 kHz. The STFT uses the Hamming window with length of 16 ms and shift of 8 ms. The CTF length is $Q = 8$ frames. The RLS forgetting factor λ is computed using $\rho = 1$. The smoothing factor β is set to 0.9. The exponentiated gradient update factor is $\eta = 0.07$. The smoothing factor η' is set to 0.065. The entropy regularization factor is $\gamma = 0.1$. For the tracker, the covariance matrix is set to be isotropic $\mathbf{\Sigma} = 0.03\mathbf{I}_2$. The threshold giving birth to a new identity is $\tau_1 = 0.75$ and $L = 3$. To decide whether a person is speaking or is silent, $L' = 3$ frames are used, with a threshold $\delta = 0.15$. At each time instance, the VEM algorithm

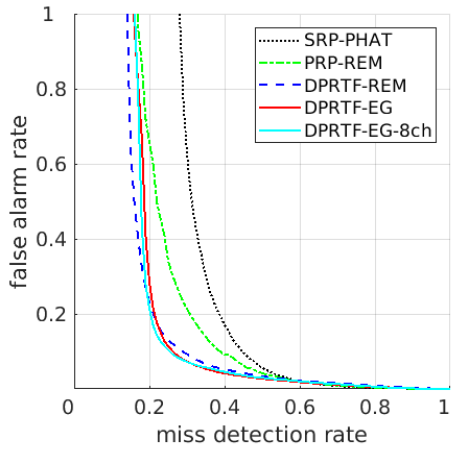


Fig. 3: ROC curve for the LOCATA dataset.

has 5 iterations. Corresponding to the STFT frame shift, i.e. 8 ms, the frame rate of the proposed system is 125 frames per second.

3) Comparison with Baseline Methods: The proposed method is evaluated both in "frame-wise localization" mode and in "tracker" mode. In the first mode, the frame-wise online localization module of Section III is applied without the tracker of Section IV. Instead, it is followed by the peak selection process described in [12]. This method is referred to as DP-RTF using EG (DPRTF-EG). In tracker mode, DPRTF-EG is directly followed by the proposed VEM tracker, without peak selection. It is then simply referred to as VEM-tracker. In that case, the directions of active speakers are given by the state variable, and the continuity of the speaker tracks is given by the assignment variable. We compare DPRTF-EG with several baseline methods:

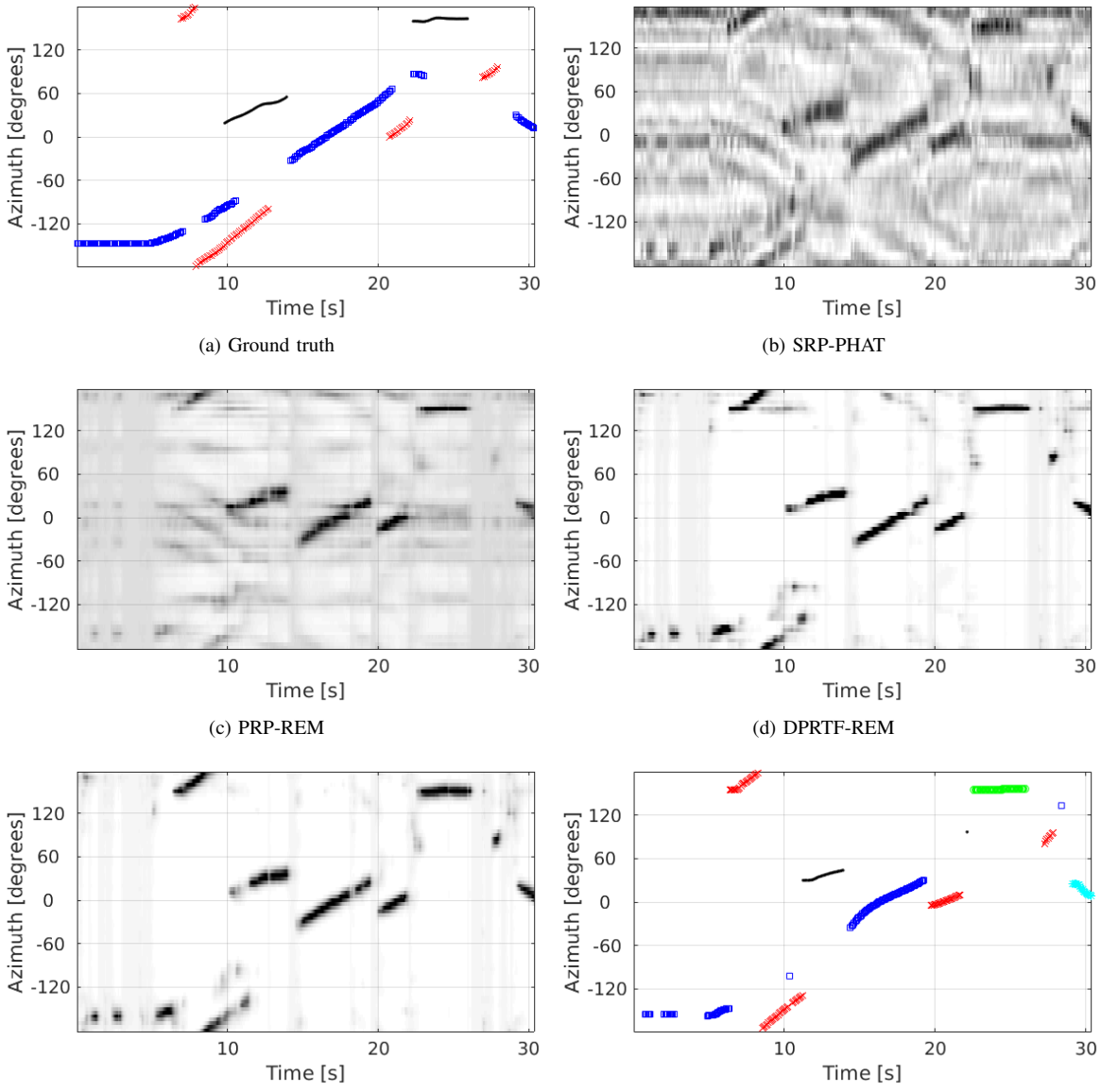
- The standard beamforming-based localization method called SRP using phase transform (PHAT) (SRP-PHAT) [3]. The same STFT configuration and candidate directions are used for SRP-PHAT and for the proposed method. The steering vector for each candidate direction is derived from the HRTFs and TDOAs for the Kinovis-MST and LOCATA datasets, respectively. The frame-wise SRP is recursively smoothed with a smoothing factor set to 0.065.
- A method combining PRP features, CGMM model and parameter update using REM [16], referred to as PRP-REM. We also combine the DPRTF features and CGMM with REM (referred to as DPRTF-REM). This is to evaluate the proposed DP-RTF feature w.r.t. PRP, and the EG-based online parameters update method w.r.t. REM. For both baselines, the STFT and CGMM settings are the same as for the proposed method. The updating factor of REM is set to 0.065.

4) Evaluation Metrics: The detected speakers should be assigned to the actual speakers for performance evaluation. This is done using a greedy matching algorithm. First the azimuth difference for all possible detected-actual speaker pairs are computed, then the detected-actual speaker pair with the smallest difference is picked out as a matched pair. This procedure is iterated until the detected or actual speakers are all picked out. For each matched pair, the detected speaker is then considered to be successfully localized if the azimuth difference is not larger than $15°$. The absolute error is calculated for the successfully localized sources. The mean absolute error (MAE) is computed by averaging the absolute error of all speakers and frames. For the unsuccessful localizations, we count the miss detection (MD) (speaker active but not detected) and false alarms (FAs) (speaker detected but not active). Then the MD and FA rates are computed, using all the frames, as the percentage of the total MDs and FAs out of the total number of actual speakers, respectively. In addition to these localization metrics, we also count the identity switches (IDs) to evaluate the tracking continuity. ID is an absolute number. It represents the number of the identity changes in the tracks for a whole test sequence.

The computation time is measured with the real-time factor (RF), which is the processing time of a method divided by the length of the processed signal. Note that all the methods are implemented in MATLAB.

B. Results for LOCATA Dataset

For convenience, both the spatial spectrum of SRP-PHAT and the CGMM component weights profile will be referred to as heatmaps. Fig. 2 shows an example of a result obtained with a LOCATA sequence. Two speakers are moving and continuously speaking with short pauses. The SRP-PHAT heatmap (Fig. 2 (b)) is cluttered due to the non ideal beampattern of the microphone array and to the influence of reverberation and noise. For most of the time, SRP-PHAT has prominent response power for the true speaker directions. Localization of the most dominant speaker can be made by selecting the direction with the largest response power. However, it is difficult to correctly count the number of active speakers and localize less dominant speakers, since there exist a number of spurious peaks. PRP-REM (Fig. 2 (c)) exhibits a clearer heatmap compared to SRP-PHAT, but there exist some spurious trajectories as well, since the PRP features are contaminated by reverberation. DPRTF-REM (Fig. 2 (d)) removes most of the spurious trajectories, which illustrates the robustness of the proposed DP-RTF feature against reverberation. From Fig. 2 (e), it can be seen that the proposed EG algorithm further removes the interferences by applying the entropy regularization. In addition, the peak evolution is smoother compared with Fig. 2 (d), which is mainly due to the use of the spatial smoothing. Fig. 2 (f) illustrates the result obtained with the proposed VEM tracker, with DPRTF-EG providing the observations. The proposed tracker gives smoother and cleaner results compared with the other methods. Even when the observations have a low weight, the tracker is still able to give the correct speaker trajectories. This is ensured by the second term in (37) which exploits the source dynamics model and continues to provide localization information even when $w_{t,d}$ (and/or α_{tdn}) becomes small. As a result, the tracker is able to preserve the identity of speakers in spite of the



(a) Ground truth

(b) SRP-PHAT

(c) PRP-REM

(d) DPRTF-REM

(e) DPRTF-EG

(f) VEM-tracking

Fig. 4: Results of speaker localization and tracking for one sequence of the Kinovis-MST dataset. (a) Ground truth trajectory and voice activity (red for speaker 1, black for speaker 2, blue for speaker 3). (b)-(e) One-dimensional heat maps as a function of time for the four tested localization methods. (f) Results for the proposed VEM-based tracker.

(short) speech pauses. In the presented sequence example, the estimated speaker identities are quite consistent with the ground truth.

To empirically evaluate the quality of the heatmaps provided by the localization methods, we computed the receiver operating characteristic (ROC) curve (MD rate versus FA rate) for the LOCATA dataset by varying the peak selection threshold, for each tested method, Fig. 3. For the ROC curve, the closer to the left-bottom the better. As already mentioned, in addition to using four microphones, we also tested an eight-microphone configuration, which is referred to as DPRTF-EG-8ch.

By analyzing the ROC curves, one notices that the methods based on DP-RTF perform better than SRP-PHAT and than PRP-REM, which is consistent with the heatmaps of Fig. 2: SRP-PHAT and PRP-REM are more sensitive to the presence of reverberations than the proposed methods. The performance of both DPRTF-REM and DPRTF-EG cannot be easily discriminated using the ROC curves. DPRTF-EG-8ch performs slightly better than DPRTF-EG, which means that the performance of the proposed method can be slightly improved by increasing the number of microphones. One may conclude that the proposed method is well suited when only a small number of microphones are available. With all methods, the FA rate can be trivially decreased to be close to 0 by increasing the peak selection threshold. However, the MD rate cannot be decreased to 0 even with a very small peak-selection threshold, since some speech frames that are actually present cannot be detected as the heatmap peaks due to the influence of noise

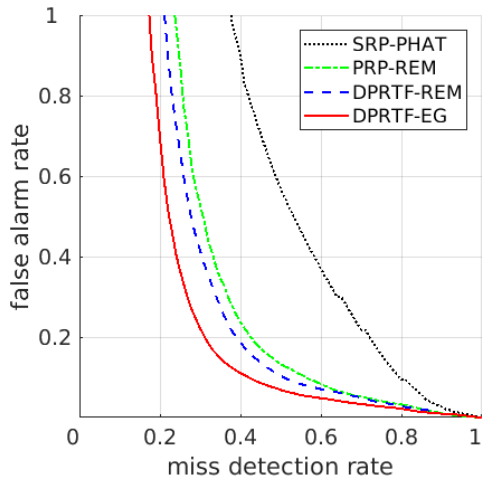


Fig. 5: ROC curve for the Kinovis-MST dataset.

and reverberation, and to a possible latency in the detection.

TABLE I: Localization and tracking results for the LOCATA data.

	MD rate (%)	FA rate (%)	MAE (°)	IDs	RF
SRP-PHAT	39.2	18.6	5.2	-	0.06
PRP-REM	30.9	19.6	5.0	-	0.30
DPRTF-REM	23.3	15.2	4.6	-	0.97
DPRTF-EG	23.9	13.0	4.0	-	0.97
DPRTF-EG-8ch	22.7	13.2	4.1	-	3.03
VEM + EG	22.7	12.4	4.1	10	2.05
VEM + EG-8ch	22.9	11.0	3.2	6	4.05

For each curve, a good balance between FA rate and MD rate is achieved at the left-bottom corner, which can be detected as the point with the minimum distance to the origin. The average localization results corresponding to this optimal left-bottom point are summarized in Table I for each tested method. It can be seen that, besides MD and FA, the DPRTF-based methods achieve smaller MAE than SRP-PHAT and PRP-REM, since the proposed DP-RTF features are robust against reverberation and thus leads to smaller biases for the heatmap peaks. DPRTF-EG has a higher MD rate than DPRTF-REM, while it also has lower FA rate, and a lower MAE, due to the effect of entropy regularization. With eight microphones, i.e. DPRTF-EG-8ch, MD is 1% smaller than the MD of DPRTF-EG, since the use of a non coplanar microphone setup provides more accurate localization than a coplanar setup. The proposed tracker performs the best in terms of MD and of FA. For the four-microphone configuration, the tracker slightly reduces FA compared to DPRTF-EG. It also reduces the MD score since some correct speaker trajectories can be recovered even when the observations have (very) low weights, as explained above. In addition, the MAE is noticeably reduced when more microphones are used by the VEM tracker, which is not the case with the DPRTF-EG localizer. This phenomenon indicates that, compared with the localizer, the tracker is able to better exploit additional information available with extra microphones, namely to revise the speaker trajectory estimation, since the state dynamics of the tracker helps correcting the possibly inaccurate additional

TABLE II: Localization and tracking results for the Kinovis-MST dataset.

	MD rate (%)	FA rate (%)	MAE (°)	IDs	RF
SRP-PHAT	60.0	37.1	5.5	-	0.07
PRP-REM	40.3	23.1	5.1	-	0.32
DPRTF-REM	37.6	22.0	5.5	-	0.73
DPRTF-EG	31.4	19.5	5.3	-	0.73
VEM + EG	31.1	11.7	4.9	11	2.12

localization information. The proposed tracker achieves quite consistent speaker ID estimation. For the whole LOCATA dataset, only ten identity switches were observed when using DPRTF-EG, and this number is reduced to six when using DPRTF-EG-8ch. The remaining identity switches are mainly due to speakers with crossing trajectories, a hard case for multiple audio-source tracking.

As for the computation time, SRP-PHAT has the smallest RF. Based on the fact that the RFs of DPRTF-REM and DPRTF-EG are identical, we can conclude that the REM algorithm and the proposed EG algorithm have comparable computational complexities. The RFs of PRP-REM, DPRTF-REM (or DPRTF-EG) and DPRTF-EG-8ch are different due to different computational complexities for feature estimation, more precisely due to the different dimensions of the vector to be estimated. The CTF identification used for DP-RTF estimation solves an RLS problem with the unknown CTF vector $\tilde{\mathbf{a}}_f \in \mathbb{C}^{(IQ-1)\times 1}$. Remind that I and Q denote the number of microphones and the CTF length, respectively. In the present work, we have set $I = 4/Q = 8$ for DPRTF-REM (or DPRTF-EG), $I = 8/Q = 8$ for DPRTF-EG-8ch. PRP is defined based on the narrow-band assumption, or equivalently based on the CTF with $Q = 1$, thence we have $I = 4/Q = 1$ for PRP-REM. The proposed localization method, i.e. DPRTF-EG with four microphones, has an RF smaller than one, which means it can be run in real time. The RF for the proposed tracker (VEM) is computed by the sum of the localization time and of the tracking time. For acoustic tracking, the tracker observes an direction of arrival (DOA) estimate every 8 ms. However, an 8 ms speaker motion is small. Thus in practice, the tracker uses one DOA estimate per 32 ms intervals, which leads to an RF of 2.05 for the four-channel (4ch) case and 4.05 for the eight-channel (8ch) case. The RF can be further improved by using less DOA estimates.

C. Results for Kinovis-MST Dataset

Fig. 4 shows an example of result for a Kinovis-MST sequence. Three participants are moving and intermittently speaking. It can be seen that, for many frames, the response power of SRP-PHAT and the CGMM component weights of PRP-REM corresponding to the true active speakers are not prominent, compared to the spurious trajectories. Again, DPRTF-REM and DPRTF-EG provide much better heatmaps, though they also miss some speaking frames, e.g. at the beginning of Speaker 3's trajectory (in blue). The possible reasons are i) the NAO robot (v5) has a relative strong ego-noise [38], and thus the signal-to-noise ratio of the recorded

signals is relative low, and ii) the speakers are moving with a varying source-to-robot distance and the direct-path speech is contaminated by more reverberations when the speakers are distant. Overall, DPRTF-REM and DPRTF-EG are able to monitor the moving, appearance, and disappearance of active speakers for most of the time, with a small time lag due to the temporal smoothing.

This kind of recording/scenario is very challenging for the tracking method, especially for speaker identity preservation, since the participants are intermittently speaking and moving. In a general manner, the proposed tracker achieves relatively good results, as illustrated in Fig. 4 (f). The tracked trajectories are smooth and clean. If the true trajectory of one speaker has an approximately constant direction, the tracker is able to re-identify the speaker even after a long silence thanks to the above-mentioned combination of observations and dynamics in (37), e.g. Speaker 1's trajectory in red. In the case that the speaker changes his/her movement when he/she is silent, the track can be lost. When the person speaks again, it is indeed difficult to re-idendify him/her based on the dynamics estimated before the silence period. The tracker may then prefer to give birth to a new speaker. This is illustrated by the black trajectory turning into green, and the blue trajectory turning into cyan in Fig. 4. Note that the silence periods are here much longer than in the LOCATA example of Fig. 2.

Fig 5 show the ROC curves for the Kinovis-MST dataset. Compared to the ROC curves for the LOCATA dataset, all the four localization methods have a worse ROC curve, especially along the MD rate axis, for the reasons mentioned above. Table II summarizes the localization and tracking results for the optimal bottom-left point of the ROC curves. It can be seen that, for the four localization methods, MAEs are quite close, namely the heatmap peaks have similar biases. Compared with the results for the LOCATA dataset, the advantage of the proposed tracker is more significant for the Kinovis-MST dataset. In particular, the FA rate is reduced by 7.8% relatively to DPRTF-EG, and is similar to the FA rate obtained with the LOCATA dataset. This means that the dynamic model associated with the tracker can efficiently reduce the influence of incorrect source localizations caused by noise and by complex source movements. The identity switches are mainly caused by speakers changing their direction of movement while during silent periods, as discussed above. Compared to the LOCATA dataset, DPRTF-EG has smaller RF, since the Kinovis-MST dataset is noisier and more noise frames are skipped in the RLS algorithm.

D. Dicussion

The experimental results obtained with the two datasets clearly show the effectiveness of the proposed method based on DP-RTF estimation, multiple speaker localization and variational tracking. To improve robustness, temporal smoothing is used, which leads to localization/tracking latency. This latency causes MD and FA observed at both the beginning and the end of continuous speech segments. However, it can be observed from the examples shown in Fig. 2 and Fig. 4 that the latency is not that severe. The Kinovis-MST dataset is more challenging than the LOCATA dataset for speaker localization/tracking, due to its lower SNR and the presence of casual speaking style. Even though, the proposed methods achieve a comparable FA rate with the two datasets. Concerning the MD score, when applied to Kinovis-MST, the method yields larger MD rates than when applied to LOCATA. This is due to the large number of TF bins dominated by a high SNR score, present in the Kinovis-MST recordings. The tracker's dynamics attenuate the influence of these TF bins to a limited extend.

VI. Conclusion

In this paper, we proposed and combined i) a recursive DP-RTF feature estimation method, ii) an online multiple-speaker localization method, and iii) an multiple-speaker tracking method. The resulting framework provides online speaker counting, localization and consistent tracking (i.e. preserving speaker identity over a track in spite of intermittent speech production). The three algorithms are computationally efficient. In particular the tracking algorithm implemented in variational Bayesian framework yields a tractable solver under the form of VEM. Experiments with two datasets, recorded in realistic environment, verify that the proposed method is robust against reverberation and noise. Moreover, the tracker is able to efficiently track multiple moving speakers, detect whether they are speech or they are silent, as long as the motion associated with silent people is smooth. However, the tracking of the person from silent to active remains a difficult task. The combination of the proposed method with speaker identification will be addressed in the future.

The proposed VEM tracker can be easily adapted to work in tandem with any frame-wise localizer providing source location estimates and/or corresponding weights (and if no weights are provided by the localizer, the tracker can be applied with all weights set to one). This makes the proposed tracker very flexible, and easily reusable by the audio processing research community.

References

[1] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[2] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on applied signal processing*, vol. 2006, pp. 170–170, 2006.

[3] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays* (M. S. Brandstein and D. Ward, eds.), pp. 157–180, Springer, 2001.

[4] C. T. Ishi, O. Chatot, H. Ishiguro, and N. Hagita, "Evaluation of a music-based real-time sound localization of multiple sound sources in real noisy environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2027–2032, 2009.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing,*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[7] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1692–1703, 2015.

[8] Y. Huang and J. Benesty, "Adaptive multichannel time delay estimation based on blind system identification for acoustic source localization," in *Adaptive Signal Processing*, pp. 227–247, Springer, 2003.

[9] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1110–1124, 2003.

[10] T. G. Dvorkind and S. Gannot, "Time difference of arrival estimation of speech source in a noisy and reverberant environment," *Signal Processing*, vol. 85, no. 1, pp. 177–204, 2005.

[11] K. Kowalczyk, E. A. Habets, W. Kellermann, and P. A. Naylor, "Blind system identification using sparse learning for TDOA estimation of room reflections," *IEEE Signal Processing Letters*, vol. 20, no. 7, pp. 653–656, 2013.

[12] X. Li, L. Girin, R. Horaud, and S. Gannot, "Multiple-speaker localization based on direct-path features and likelihood maximization with spatial sparsity regularization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1997–2012, 2017.

[13] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1305–1319, 2007.

[14] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 546–555, 2009.

[15] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[16] O. Schwartz and S. Gannot, "Speaker tracking using recursive EM algorithms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 392–402, 2014.

[17] N. Roman and D. Wang, "Binaural tracking of multiple moving sources," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 728–739, 2008.

[18] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer, and B. Rafaely, "Bearing-only acoustic tracking of moving speakers for robot audition," in *IEEE International Conference on Digital Signal Processing (DSP)*, pp. 1206–1210, 2015.

[19] Y. Ban, L. Girin, X. Alameda-Pineda, and R. Horaud, "Exploiting the complementarity of audio and visual data in multi-speaker tracking," in *ICCV Workshop on Computer Vision for Audio-Visual Media*, vol. 3, 2017.

[20] Z. Liang, X. Ma, and X. Dai, "Robust tracking of moving sound source using multiple model kalman filter," *Applied acoustics*, vol. 69, no. 12, pp. 1350–1355, 2008.

[21] J. Vermaak and A. Blake, "Nonlinear filtering for speaker tracking in noisy and reverberant environments," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 5, pp. 3021–3024, IEEE, 2001.

[22] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[23] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal Bayesian fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[24] M. F. Fallon and S. J. Godsill, "Acoustic source localization and tracking of a time-varying number of speakers," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1409–1415, 2012.

[25] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems*, vol. 55, no. 3, pp. 216–228, 2007.

[26] V. Cevher, R. Velmurugan, and J. H. McClellan, "Acoustic multitarget tracking using direction-of-arrival batches," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2810–2825, 2007.

[27] B.-N. Vo, S. Singh, and W. K. Ma, "Tracking multiple speakers using random sets," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 2, pp. ii–357, IEEE, 2004.

[28] W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using tdoa measurements: A random finite set approach," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

[29] B.-N. Vo and W.-K. Ma, "The gaussian mixture probability hypothesis density filter," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4091–4104, 2006.

[30] X. Li, B. Mourgue, L. Girin, S. Gannot, and R. Horaud, "Online localization of multiple moving speakers in reverberant environments," in *The Tenth IEEE Workshop on Sensor Array and Multichannel Signal Processing*, 2018.

[31] J. Kivinen and M. K. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.

[32] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *IEEE Transactions on signal processing*, vol. 43, no. 12, pp. 2982–2993, 1995.

[33] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of relative transfer function in the presence of stationary noise based on segmental power spectral density matrix subtraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 320–324, 2015.

[34] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural computation*, vol. 15, no. 4, pp. 915–936, 2003.

[35] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "Em algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 12, pp. 2402–2415, 2016.

[36] H. W. Löllmann, C. Evers, A. Schmidt, H. Mellmann, H. Barfuss, P. A. Naylor, and W. Kellermann, "The LOCATA challenge data corpus for acoustic source localization and tracking," in *IEEE Sensor Array and Multichannel Signal Processing Workshop*, (Sheffield, UK), July 2018.

[37] X. Li, R. Horaud, L. Girin, and S. Gannot, "Voice activity detection based on statistical likelihood ratio with adaptive thresholding," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 1–5, 2016.

[38] X. Li, L. Girin, F. Badeig, and R. Horaud, "Reverberant sound localization with a robot head based on direct-path relative transfer function," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2819–2826, IEEE, 2016.